Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 75-78 recibido 21-11-2017 revisado 07-01-2018 aceptado 07-02-2018

Interfaces de Lenguaje Natural para la Consulta y Recuperación de Información de Bases de Conocimiento Basadas en Ontologías

Natural Language Interfaces for Querying and Retrieving Information from Ontology-based Knowledge Bases

Mario Andrés Paredes Valverde
Universidad de Murcia
Facultad de Informática Campus Espinardo
Espinardo, 30100, Murcia, España
marioandres.paredes@um.es

Resumen: Tesis doctoral titulada "Interfaces de lenguaje natural para la consulta y recuperación de información de bases de conocimiento basadas en ontologías ", defendida por Mario Andrés Paredes Valverde en la Universidad de Murcia y elaborada bajo la dirección de los doctores Rafael Valencia García (Universidad de Murcia) y Miguel Ángel Rodríguez García (King Abdullah University of Science & Technology). La defensa tuvo lugar el 23 de mayo de 2017 ante el tribunal formado por los doctores Juan Miguel Gómez Berbís (Presidente, Universidad Carlos III de Madrid), Francisco García Sánchez (Secretario, Universidad de Murcia) y la doctora Catalina Martínez Costa (Vocal, Medical University of Graz) y la tesis obtuvo la mención Cum Laude y Doctor Internacional.
Palabras clave: Procesamiento de lenguaje natural, web semántica, linked data

Abstract: Ph.D. thesis entitled "Natural language interfaces for querying and retrieving information from ontology-based knowledge bases" written by Mario Andrés Paredes Valverde at the University of Murcia under the supervision of the Ph.D. Rafael Valencia García (University of Murcia) and Ph.D. Miguel Ángel Rodríguez García (King Abdullah University of Science & Technology). The viva voice was held on the 23rd May 2017 and the members of the commission were the Ph.D. Juan Miguel Gómez Berbís (President, University Carlos III of Madrid), Ph.D. Francisco García Sánchez (Secretary, University of Murcia) and Ph.D. Catalina Martínez Costa (Vocal, University of Graz) and the thesis obtained the mention Cum Laude and International Doctor.
Keywords: Natural language processing, semantic web, linked data

1 Introducción

El exponencial crecimiento de información disponible en la web e intranets ha dado paso a la necesidad de contar con mecanismos capaces de procesar y comprender dicha información y con ello resolver necesidades específicas. Ante esta situación surge la web semántica, la cual, de acuerdo con Berners-Lee et al. (2001) añade a la información de la web actual una estructura bien definida a través de un conjunto de atributos, valores y relaciones, para lo cual emplea una de las tecnologías más sobresalientes de su arquitectura, que son las ontologías. Diversos individuos y organizaciones de dominios tales como las finanzas (Salas-Zárate et al., 2016) y servicios en la nube (Rodríguez-García et al., 2014) han adoptado las ontologías para publicar su información. Sin embargo, uno de los enfoques más extendidos para el acceso a esta información es el lenguaje formal de consulta SPARQL cuyo uso demanda un alto nivel de conocimiento en tecnologías como RDF y expresiones de lenguaje de consulta, así como el conocimiento previo de la estructura de datos de la base de conocimiento subyacente.

Ante estos hechos, existe la necesidad de hacer accesible la información de la web semántica a todo tipo de usuarios, sean expertos u ocasionales. De acuerdo con Cimiano et al. (2008) el paradigma de recuperación de información basado en lenguaje natural es

ISSN 1135-5948. DOI 10.26342/2018-60-9 © 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

generalmente considerado como el más intuitivo desde un punto de vista de uso, pues oculta al usuario la formalidad de una base de conocimientos basada en ontologías, así como el lenguaje de consulta ejecutable, permitiendo a los usuarios emplear todo el poder comunicativo del lenguaje natural en lugar de verse forzados a utilizar un lenguaje limitado.

En esta tesis doctoral se propone una solución basada en lenguaje natural y ontologías para la consulta y recuperación de información de bases de conocimiento. La solución propuesta aprovecha la tecnología de la web semántica de dos maneras. La primera de ellas consiste en procesar la ontología de la base de conocimiento para generar un vocabulario que le permita conocer los términos comúnmente utilizados por los usuarios en el dominio modelado, y de esta manera poder relacionar los elementos contenidos en la pregunta del usuario con aquellos descritos en la base de conocimiento. La segunda, consiste en utilizar un modelo ontológico independiente del dominio para representar tanto la estructura sintáctica de la pregunta, como el contexto de esta en términos de la base de conocimiento. Para obtener tal representación, se aplican técnicas de procesamiento de lenguaje natural (PLN), entre las que destaca el análisis de dependencias. A través de esta técnica se obtiene una representación sintáctica de la pregunta que guarda una estrecha relación con las tripletas RDF que forman el patrón de grafos a ser obtenido de la base de conocimiento. Este hecho ayuda en gran medida a generar las consultas SPARQL respectivas con base en un conjunto de plantillas de consulta independientes del dominio. A continuación, se describe de manera general el trabajo de investigación doctoral.

2 Objetivos

El objetivo principal de esta tesis es desarrollar soluciones basadas en tecnologías de procesamiento de lenguaje natural y web semántica que permitan reducir la brecha existente entre el usuario y las bases de conocimiento a través del lenguaje natural. Con respecto a los objetivos específicos de la tesis, estos se resumen de la siguiente manera:

1. Diseño e implementación de un modelo ontológico independiente del dominio para la representación de la estructura sintáctica y contexto de la pregunta en lenguaje natural.
2. Diseño de la arquitectura de una interfaz de lenguaje natural para bases de conocimiento basadas en ontologías.
3. Diseño e implementación de un proceso de análisis de preguntas basado en técnicas de procesamiento de lenguaje natural y web semántica.
4. Diseño e implementación de un proceso de generación de consultas SPARQL a partir de una representación semántica de la pregunta en lenguaje natural.
5. Validación de los resultados obtenidos por medio de bases de conocimiento basadas en Linked Data.

3 Estructura de la tesis

La tesis se ha organizado en 6 capítulos que se describen a continuación.

Capítulo 1. Este capítulo provee una breve introducción a las motivaciones del trabajo de investigación y a la metodología seguida para cumplir con los objetivos establecidos.

Capítulo 2. Esta sección proporciona una descripción del estado actual de las tecnologías involucradas en la investigación, que son web semántica, PLN e interfaces de lenguaje natural.

Capítulo 3. Este capítulo discute a detalle la principal motivación para llevar a cabo la investigación. Además, provee tanto el objetivo general como los objetivos específicos establecidos. Finalmente, describe la metodología seguida en esta investigación.

Capítulo 4. Este capítulo describe la arquitectura y funcionamiento general de la interfaz de lenguaje natural para bases de conocimiento basadas en ontologías propuesta en la tesis. También, describe el modelo ontológico de la pregunta que permite describir su estructura y contexto.

Capítulo 5. Este capítulo describe los experimentos de evaluación realizados para medir la efectividad de la interfaz de lenguaje natural, la cual se base en su capacidad de proveer la respuesta correcta a una pregunta en lenguaje natural a partir de una base de conocimientos. Estos experimentos se llevaron a cabo en dos bases de conocimiento con el objetivo adicional de comprobar la portabilidad de la interfaz.

Capítulo 6. Este apartado describe las conclusiones, y discute las principales contribuciones y limitaciones del trabajo

realizado, así como las posibles vías futura que permitan direccionarlas.

4 Contribuciones

Las principales contribuciones de esta tesis doctoral se resumen a continuación.

Modelo ontológico de la pregunta. Modelo ontológico que permite describir la estructura sintáctica de la pregunta, así como el contexto de esta en términos de la base de conocimiento del dominio y de las relaciones existentes entre ellos. La obtención de la estructura sintáctica de la pregunta se basa en la técnica de análisis de dependencias. Esta técnica obtiene relaciones binarias entre los elementos de la pregunta, las cuales, gracias al modelo de la pregunta, pueden ser representadas en forma de tripletas sujeto-predicado-objeto.

Adaptación de una clasificación de preguntas y respuestas al contexto de las bases de conocimiento basadas en ontologías. Esta contribución consiste en adaptar la clasificación de preguntas propuesta por Moldovan et al. (2000). Esta adaptación consistió en sustituir los tipos de respuesta esperados por clases establecidas en ontologías y vocabularios que han sido ampliamente adoptados por individuos y organizaciones para representar su información. Gracias a este proceso, es posible delimitar el espacio de búsqueda de tal forma que los recursos a obtener deberán ser solo aquellos que correspondan con el tipo de datos establecido, o sean subclase de este.

Conjunto de plantillas de tripletas RDF. Esta contribución consiste en un conjunto de plantillas de tripletas RDF las cuales corresponden a las relaciones semánticas existentes entre los elementos de interés identificados en la pregunta que son representados mediante el modelo ontológico de la pregunta propuesta en esta tesis. El conjunto de plantillas ha probado ser independiente del dominio y permite la generación de consultas SPARQL formadas por múltiples tripletas.

Validación de la interfaz en diferentes dominios. El proceso de validación de la interfaz se llevó a cabo en dos dominios bien diferenciados, a saber, DBPedia y MusicBrainz, y cuyos resultados se publicaron en Paredes-Valverde et al. (2015) y Paredes-Valverde et al. (2016) respectivamente. Los experimentos realizados involucraron corpus de preguntas en lenguaje natural utilizados por la comunidad científica para evaluar interfaces de lenguaje natural orientadas a fuentes de datos semánticas, y un conjunto de preguntas en lenguaje natural elaboradas por usuarios potenciales ajenos al trabajo de investigación. Los resultados obtenidos en ambos dominios no varían significativamente uno del otro, lo cual se puede interpretar como un buen nivel de portabilidad por parte de la interfaz desarrollada.

5 Limitaciones

A pesar de que los resultados de evaluación obtenidos por la interfaz propuesta en esta tesis doctoral lucen alentadores, somos conscientes que este enfoque tiene ciertas limitaciones que, sin embargo, pueden ser direccionadas a futuro. Estas limitaciones se describen a continuación.

Tipos de pregunta soportadas. La interfaz de lenguaje natural permite el uso de preguntas factuales, es decir, aquellas que esperan como respuesta un hecho concreto. Por ejemplo, el nombre de una persona o lugar, la altura de una persona, entre otros. Además, la interfaz permite el uso de oraciones imperativas para solicitar información. En este sentido, es importante considerar más tipos de pregunta como de opción múltiple, verdadero/falso, entre otras. Para direccionar esta limitación, se planea el análisis de un corpus de preguntas de este tipo, que nos permita identificar las relaciones de dependencia que ayudarían a obtener una representación semántica de la pregunta.

Problemas de ambigüedad. La ambigüedad se refiere al fenómeno que se presenta cuando una palabra, un sintagma, o una oración puede ser interpretada de más de una forma. A pesar de que la interfaz propuesta implementa mecanismos para afrontar algunos casos de ambigüedad, somos conscientes de que estos presentan limitaciones que les impiden direccionar de mejor manera este problema. Para hacer frente a esta limitación, se plantea la integración de mecanismos de retroalimentación que permitan al usuario desambiguar las preguntas, ya sea a través de la reformulación de la pregunta o a través de seleccionar alguna opción de un conjunto provisto por la interfaz.

Nombre de individuos compuestos por múltiples palabras. La interfaz desarrollada identifica dentro de la pregunta aquellos elementos que hagan referencia a un individuo de la base de conocimiento, como lo puede ser

un libro o una canción. Sin embargo, en ocasiones el nombre del individuo está compuesto por múltiples palabras. Cuando este fenómeno ocurre, la interfaz puede reconocer este nombre, siempre y cuando combine el uso de mayúsculas y minúsculas. Este fenómeno representa un gran reto en el contexto de PLN, tal como se describe en Sag et al. (2002), donde de igual manera se presentan técnicas que podrían ser implementadas en esta interfaz, tal como el uso de reglas o métodos estadísticos.

6 Trabajo a futuro

A continuación, se describen temas que no han sido abordados por la interfaz propuesta y que proporcionan nuevas líneas de investigación.

Conjuntos de datos distribuidos y enlazados. Existe un gran número de individuos y organizaciones que han adoptado ya el enfoque de la web semántica, y en específico, el de Linked Data, para publicar sus datos. Debido a esto, es importante considerar esa distribución al momento de buscar una respuesta, pues en ocasiones esta puede depender de más de una base de conocimientos. En esta línea de investigación, se propone analizar el estado del arte sobre la consulta y recuperación de información de fuentes de información descentralizadas como Linked Data. Algunos de los enfoques más sobresalientes son las consultas federadas a SPARQL endpoints, fragmentos de patrones de tripletas y flujos de Linked Data. Tras el estudio podremos establecer un punto de partida para abordar el problema en cuestión e integrarlo en la interfaz de esta tesis.

Multilingüismo. Actualmente, la mayoría de las interfaces de lenguaje natural orientadas a bases de conocimiento basadas en ontologías no son capaces de responder a preguntas formuladas en múltiples lenguas. Esta tarea requiere que los recursos descritos en las ontologías (clases, propiedades e individuos) cuenten con una propiedad a través de la cual referenciarlos en cada una de las lenguas a considerar. Dicho esto, esta línea de investigación propone adaptar la interfaz propuesta en esta tesis a otra lengua, concretamente, al español. Esto demandará analizar herramientas para el análisis sintáctico de dependencias, y así reducir aún más la brecha existente entre usuarios y bases de conocimiento basadas en ontologías.

Agradecimientos

Mario Andrés Paredes Valverde es apoyado por la Comisión Nacional de Ciencia y Tecnología (CONACyT) y la Secretaria de Educación Pública (SEP).

Bibliografía

Berners-Lee, T., J. Hendler, O. Lassila. 2001. The Semantic Web. *Scientific American*. 284 (5): 28-37.

Cimiano, P., P. Haase, J. Heizmann, M. Mantel, y R. Studer. 2008. Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system. *Data & Knowledge Engineering*. 65 (2): 325-354.

Moldovan, D., S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum, y V. Rus. 2000. The structure and performance of an open-domain question answering system. En *Proceedings of the 38th annual meeting on association for computational linguistics*, páginas 563-570.

Paredes-Valverde, M. A., M. Rodríguez-García, A. Ruiz-Martínez, R. Valencia-García, y G. Alor-Hernández. 2015. ONLI: An Ontology-Based System for Querying DBpedia Using Natural Language Paradigm. *Expert Systems with Applications.* 42 (12): 5163–76.

Paredes-Valverde, M. A., R. Valencia-García, M. Rodríguez-García, R. Colomo-Palacios, y G. Alor-Hernández. 2016. A Semantic-Based Approach for Querying Linked Data Using Natural Language. *Journal of Information Science*. 42 (6): 851–62.

Rodríguez-García, M., R. Valencia-García, F. García-Sánchez, y J. Samper-Zapater. 2014. Ontology-based annotation and retrieval of services in the Cloud. *Knowledge-based systems*. 56:15-25.

Sag, I., T. Baldwin, F. Bond, A. Copestake, y D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. En *Computational linguistic and intelligent text processing*, 1-15. Lecture notes in computer science. Springer Berlin Heidelberg.

Salas-Zárate, M., R. Valencia-García, A. Ruíz-Martínez, y R. Colomo-Palacios, 2016. Feature-based opinion mining in financial news: An ontology-driven approach. *Journal of Information Science*.